

2017/12/25-2017/12/31周报

DONE

- 《人工智能引论》：课程论文撰写和presentation准备，人工智能引论课程要求阅读ICCV等人工智能相关会议/期刊的论文并撰写报告并进行presentation。本周阅读了五篇手势识别相关的论文，基本上都与卷积神经网络息息相关。据说罗凌志老师前些天发了一篇关于用CNN对graph进行分类的文章，下周会找他聊一聊我们的想法，用CNN来跑图数据。
- 阅读论文VIGOR: [VIGOR-Interactive Visual Exploration of Graph Query Results](#)，之后投稿应该会参考这篇论文来进行。

TODO

- 新投稿：下周继续跟郭博 东明 小涛继续交流想法，希望能有一个明确的方向。
- 软件注册：画流程图，重新截图
- 芯片调研：下周会继续调研芯片

任务	截止日期	当前进度
RCAnalyzer文章投稿		已完成
大图可视化调研		开始调研芯片
关于palantir软件注册撰写		律师回复截图不够清晰，需要在高分辨率屏幕上进行截图、需要流程图。
硕士论文	春节前	已经完成开题报告，还差论文主体部分
新投稿思路确定	12月底	推翻了已有思路 要重新考虑方向

第一人称视角的手姿势估计

1. 背景介绍

由于手在日常的人类活动中，起着至关重要的作用，估计完整的手的三维姿态越来越重要。在很多场景下，比如运动控制，人机交互，虚拟/增强现实，对手姿态的估计需要在一些混乱、有干扰的环境下进行。由于最近卷积神经网络的发展，目前静态、第三人称视角、在无干扰、无遮挡环境下对手部的追踪和手势估计已经很有效。但很明显，这种场景设定在一些实际场景中并不常见。

在现实世界场景下，经常需要从第一视角来进行手势追踪估计，而且背景往往杂乱无章，手经常在跟物体进行交互时存在遮挡，而需要交互的物体形状不定，这就为手势估计和重建构成了很多具有挑战性的任务。总结而言，第一人称视角下，进行手势跟踪和估计具有如下一些挑战：

- 存在遮挡
- 背景杂乱无章（噪声）
- 第一人称视角带来的视野限制（因为相机往往放置在肩部，如图 1）
- 手和物体的交互
- 该场景下，标注数据的缺失



图 1 第一人称视角的手势追踪

本文总结了近几年在第一人称手势估计方面的顶级工作。这些方法基本上都是基于卷积神经网络进行，利用已有的，或者自己采集的，真实的或者合成的数据进行训练。一般针对与物体交互的场景进行，需要在有遮挡、杂乱无章的环境下进行。

2. 手部姿势估计的一般过程

一般而言，第一人称视角下，手部姿势估计过程可以分成两步，手部定位和姿势估计。近几年，这两者基本上都可以通过卷积神经网络实现。

手部定位：通过训练好的，多层卷积神经网络，我们可以得到一个关于手部中心位置的置信度分布图，从而可以绘制成热力图的形式，如图 2 a。

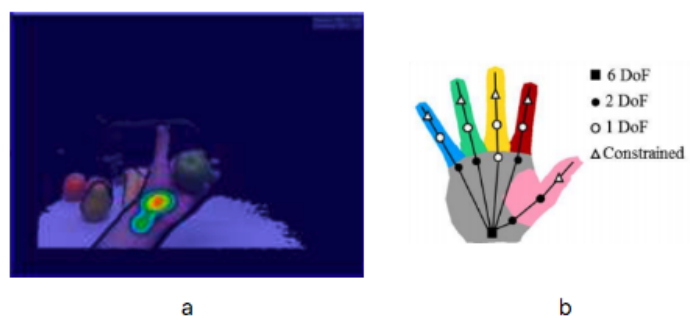


图 2 卷积神经网络输出的热力图

姿势估计：对手部姿势估计，一般都会采用 21 手关节参数[i]来估计整个手的姿态。通过手关节的角度，相对位置等信息，可以对整个手姿态进行一个重建，如图 2 b。

3. 带标注的训练数据的获取

一般而言，根据数据获取方式，可以分成三大类，(i) 完全合成的数据，(ii) 半合成数据以及(iii)真实数据。这些数据大部分是基于 RGB-D 相机获取的，也有搭配手部传感器进行采集。

1 完全合成数据

因为手动模拟手-物体的交互是一项耗时的任务，故而采用自动的，完全用计算机进行模拟的方法来合成数据。

Choi 等人的工作[ii]中，提出了一种用模型进行拟合的方式，来优化模拟手对于模拟物体的抓握姿势的方法。该工作使用了粒子群优化方法 (particle swarm optimization)，对虚拟的三维手模型和被该手握住的模型的距离误差进行最小化。之后，作者通过碰撞检测技术来判断这个虚拟手部的抓取是否有效，从而排除无效的抓取。之后作者从模型中导出了相应的手的关节角度参数。最后作者将虚拟手部的深度图插入到杂乱无章的背景中，以模仿现实世界的真是噪音的存在。

2 半合成数据

为了在真实性和数据的多样性、易获取性上做出权衡，Mueller 等人的工作 [iii] 则采用了半合成数据的方法，作者称为混合现实（merged reality）。如图 3 所示，作者用一个无标记跟踪摄像头，用以从第三视角，跟踪一个无遮挡的真实手部，产生跟踪数据（这种方法已经很成熟），从而用以操作虚拟 3D 模型手来抓取虚拟的物体。通过这个方法可以增加数据的真实性。

这种方式的优点在于，既保留了一定的手部运动真实性，又能模拟出大量不同的虚拟手部（比如肤色的不同，手指长短，体毛浓密等等），与各种不同类型物体的交互以及各种不同的场景。

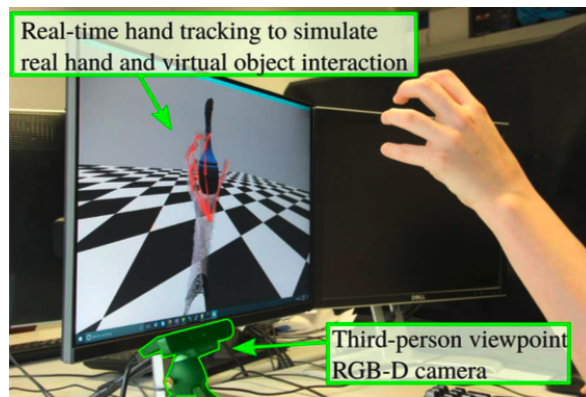


图 3 混合现实（merged reality）方式获取手部数据

3 真实数据

在 Garcia-Hernando [iv] 的工作中，采集了超过 100,000 个有 3D 手姿势注释的 RGB-D 帧，使用装在指尖上的六个磁性传感器进行数据采集，如图 4。其中包括 45 个类别，在 3 个场景中，与 25 个不同的物体进行交互。作者静心设计了不同的手部动作，以保证能覆盖较多的姿势，交互时间和运动状态。



图 4 使用传感器和 RGB-D 相机 采集第一人称数据

4. 手部定位方法 (Localization)

因为卷积神经网络 (CNN) 方法的成熟，现如今，手部定位基本上都采用卷积神经网络来进行，Choi 等人的工作，也证明了卷积神经网络相对于随机森林方法的优势[1]。在经过训练好的深度神经网络的估计之后，一般会输出一张关于手部中心位置置信度的热力图。

比如，在 Choi 等人的工作中[11]，训练了一个带有六个卷积层，以及一个非线性判定层的卷积神经网络，用以判定输入图中的手心位置以及物体位置，如图。

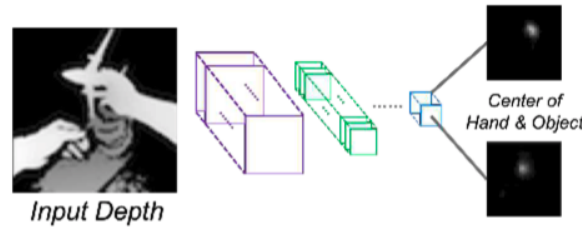


图 5 由五层卷积层与一层非线性层构成的位置估计卷积网络

Mueller 等人的工作[12]，在深度神经网络训练结果上，加入了后处理 (post processing) 过程，目的是增加数据的稳定性，因为手部中心位置不会随时间的变化而发生较大的变化)。作者保留了同一组数据，多帧图像的历史定位记录，并对每一帧图像的手部位置的定位结果做一个是否可信的判定。如果帧 t 的热力图的最大置信度小于 0.1，且出现的位置和上一次最大值位置距离大于 30，则认为其不可信，然后需要对这个最大值点进行更新：

$$\phi_t = \phi_{\{t-1\}} + \delta^k \frac{\phi_{\{c-1\}} - \phi_{\{c-2\}}}{\|\phi_{\{c-1\}} - \phi_{\{c-2\}}\|}$$

其中 $\phi_t = \phi(H_R^t)$ 是帧 t 处的更新之后的最大值位置，是 ϕ_{c-1} 上一个可置信 (confident) 的最大值位置， k 是自上个可置信的最大值起经过的帧数， δ 是逐渐对不可信的最大值进行减权 (downweight) 的衰减因子。

经过更新，最大值位置不会随着时间变化而发生较剧烈的变化。

5. 手势估计方法 (Hand Pose Estimation)

目前有的手势估计方法基本上可以分成两类，基于分类的方法和基于回归的方法。

分类方法：Rogez 等人在 2014 年发表的文章[13]中，使用了分层量化的分类

器。首先，需要构建一棵姿态类别树，每个节点都代表了一种姿态，层次越高的姿态，普遍性越高。而越接近叶节点，代表其手势的细节越多，更特殊，如图 6 所示。接下去，可以使用宽度优先搜索（BFS）来对输入的手势进行评估，对树的每一层，都可以剪掉那些评级为 0 的节点（也就是没有吻合性的姿态）。

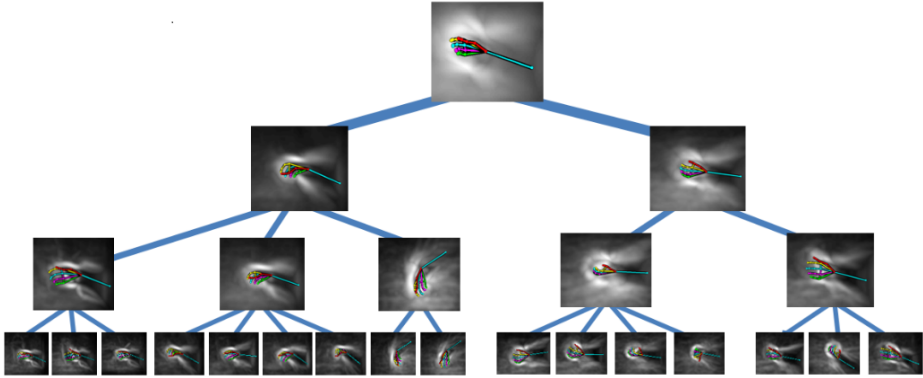


图 6 手势层次量化分类器

Choi 等人的工作[iii]，用一个卷积神经网络，训练了一个分类器，已分类手部姿态，该工作的场景中，用户会抓取一个未知的物体。在训练分类器之前，作者认为，在追踪用户的手部姿态的时候，会因为传感器噪声等，导致输出的 RGB-D 图像存在一定缺失。故而在训练该分类器的同时，训练了输入数据再生成网络，通过自动编码器的方法实现输入数据复原。自动编码器由编码器（将高维数据映射到较低的维特征空间来降低输入的维数）和解码器（通过将学到的表示，映射回高维空间来恢复原始输入）构成，两者都是卷积神经网络，作者为两者都加入了四层隐含层，如图 7。之后，作者认为手的姿态和物体形状高度相关，所以在训练手部姿态分类器的同时，也加入了另一个神经网络作为物体形状的分类器，并共享这两者的决策层，来协作学习手和物体的成对的卷积特征，以增加分类准确性，如图 7。

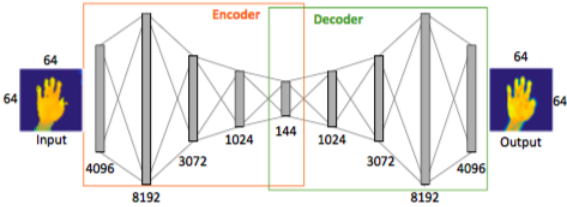


图 7 RGB-D 图 数据再生成网络

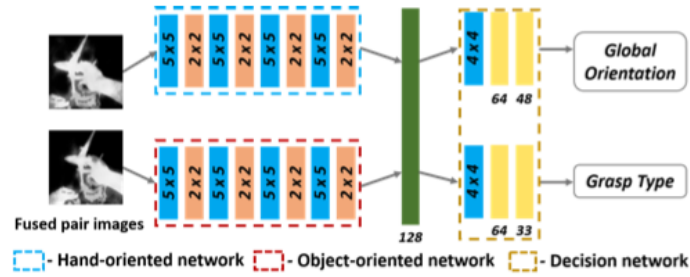


图 8 手势分类器神经网络的架构

回归方法：在 Mueller[iii]等人的工作中，则是使用了卷积神经网络，对手部姿态向量进行了回归。作者采用了一个 26 自由度的手骨骼模型，其包括 6 个用于全局平移和旋转的角度，以及 20 个关节角度，存储在向量 Θ 中，然后用一个卷积神经网络对这个向量进行了回归。Mueller 等人在这个基础上，还加入了一些约束（比如关节旋转角度的范围，手关节运动速度等），以保证回归结果的有效性。

$$\mathcal{E}(\Theta) = E_{\text{data}}(\Theta, p^G, H) + E_{\text{reg}}(\Theta)$$

其中 E_{data} 约束了相对位置， E_{reg} 约束了关节旋转角度和运动速度。

6. 引用

-
- [i] Choi C, Sinha A, Hee Choi J, et al. A collaborative filtering approach to real-time hand pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 2336-2344.
 - [ii] Choi, Chiho, et al. "Robust Hand Pose Estimation during the Interaction with an Unknown Object." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
 - [iii] Mueller, Franziska, et al. "Real-time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor." arXiv preprint arXiv:1704.02201 (2017).
 - [iv] Garcia-Hernando, Guillermo, et al. "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations." arXiv preprint arXiv:1704.02463 (2017).
 - [v] Rogez, Grégory, et al. "3d hand pose detection in egocentric RGB-D images." arXiv preprint arXiv:1412.0065 (2014).